

## Journal of Biopharmaceutical Statistics

Publication details, including instructions for authors and subscription information:

<http://www.tandfonline.com/loi/lbps20>

### A Clustering Algorithm for Multivariate Longitudinal Data

Liesbeth Bruckers<sup>a</sup>, Geert Molenberghs<sup>ab</sup>, Pim Drinkenburg<sup>c</sup> & Helena Geys<sup>ac</sup>

<sup>a</sup> I-BioStat, Universiteit Hasselt, Agoralaan, B3590 Diepenbeek, Belgium.

<sup>b</sup> I-BioStat, Katholieke Universiteit Leuven, Kapucijnenvoer 35, B3000 Leuven, Belgium.

<sup>c</sup> Janssen Research & Development, a division of Janssen Pharmaceutica NV, Beerse, Belgium.

Accepted author version posted online: 26 May 2015.



[Click for updates](#)

To cite this article: Liesbeth Bruckers, Geert Molenberghs, Pim Drinkenburg & Helena Geys (2015): A Clustering Algorithm for Multivariate Longitudinal Data, Journal of Biopharmaceutical Statistics, DOI: [10.1080/10543406.2015.1052476](https://doi.org/10.1080/10543406.2015.1052476)

To link to this article: <http://dx.doi.org/10.1080/10543406.2015.1052476>

Disclaimer: This is a version of an unedited manuscript that has been accepted for publication. As a service to authors and researchers we are providing this version of the accepted manuscript (AM). Copyediting, typesetting, and review of the resulting proof will be undertaken on this manuscript before final publication of the Version of Record (VoR). During production and pre-press, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal relate to this version also.

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the "Content") contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms & Conditions of access and use can be found at <http://www.tandfonline.com/page/terms-and-conditions>

# A Clustering Algorithm for Multivariate Longitudinal Data

Liesbeth Bruckers\*,<sup>1</sup>

Geert Molenberghs<sup>1,2</sup>

Pim Drinkenburg<sup>3</sup>

Helena Geys<sup>1,3</sup>

<sup>1</sup>I-BioStat, Universiteit Hasselt, Agoralaan, B3590 Diepenbeek, Belgium.

<sup>2</sup>I-BioStat, Katholieke Universiteit Leuven, Kapucijnenvoer 35, B3000 Leuven, Belgium.

<sup>3</sup>Janssen Research & Development, a division of Janssen Pharmaceutica NV, Beerse, Belgium.

\**E-mail*: liesbeth.bruckers@uhasselt.be

## Acknowledgements

The authors gratefully acknowledge support from IAP research Network P7/06 of the Belgian government (Belgian Science Policy) and Janssen Pharmaceutica for providing the EEG data that was used for illustrative purposes.

## Abstract

Latent growth modeling approaches, such as growth mixture models, are used to identify meaningful groups or classes of individuals in a larger heterogeneous population. But when applied to multivariate repeated measures computational problems are likely, due to the high

dimension of the joint distribution of the random effects in these mixed-effects models. This paper proposes a cluster algorithm for multivariate repeated data, using pseudo-likelihood and ideas based on  $k$ -means clustering, to reveal homogenous subgroups. The algorithm was demonstrated on an EEG data set quantifying the effect of psychoactive compounds on the brain activity in rats.

*Keywords: Cluster analysis; Multivariate longitudinal data; Joint models; EEG data.*

## **1 Introduction**

Cluster analysis (Johnson & Wichern, 2007) refers to a collection of procedures that attempt to determine natural groupings (or clusters) of observations in a population. Some refer to these techniques by ‘classification’; but this term is used by others to mean discriminant analysis (Johnson & Wichern, 2007). Discriminant analysis is related to but differs from cluster analysis. To avoid confusion, we will use ‘cluster analysis’ when referring to finding (unknown) groups in data such that observations in the same cluster are similar in some sense.

A lot of methodological work has been done to extend cluster analysis such that it can cope with complex data structures, in particular repeated measures (Nagin, 2005; Nagin and Tremblay, 2001; Putter et al., 2008; Roy and Lin, 2000). These methods perform well when studying one repeated process. When clustering based on a set of multivariate longitudinal outcomes is of interest, these methods run in computational problems. In this paper, we propose a clustering technique for multivariate longitudinal data, by fitting joint mixed-effects models and borrowing ideas from  $k$ -means clustering.

This article is organized as follows. Clustering methods for repeated data are briefly reviewed in Section 2. Joint models for longitudinal data are introduced in Section 3. The clustering algorithm for multivariate repeated data is described in Section 4 and demonstrated on EEG data in Section 5. Section 6 investigates the performance of the algorithm by means of a simulation study. Finally, Section 7 contains a discussion.

## 2 Clustering Methods for Repeated Data

Over the last decades, a lot of methodological work has been done to extend model-based cluster analysis such that it can cope with complex data structures, in particular repeated measures.

When analyzing repeated measurements data, individual differences in evolution are generally captured by random effects, often via linear mixed models (Laird and Ware, 1982; Verbeke and Molenberghs, 2000). Such random effects reflect continuous variation across individuals in growth features, such as initial stage and rate of change.

Individual differences can also be described by latent trajectory classes. The group-based approach, developed by Nagin (Nagin and Land, 1993; Land and Nagin, 1996; Nagin, 1999; Nagin and Tremblay, 2001), uses a multinomial modeling strategy to identify relatively homogeneous clusters of trajectories. The assumption is made that, conditional on group membership, the repeated measurements over time are independent. In other words, all correlation between the repeated measures is captured by the latent class.

Random effects are continuous latent variables whereas latent trajectory classes are categorical latent variables. Growth mixture modeling (GMM) uses both types of latent variables to

represent individual differences in development (Muthén and Shedden, 1999; Muthén and Muthén 1998, 2007). GMM specifies a separate latent growth model for each of the  $K$  latent subpopulations simultaneously.

Denote for a univariate longitudinally measured outcome  $y_{il}$  the  $l^{th}$  measurement available for the  $i^{th}$  subject ( $i = 1, \dots, N$ ;  $l = 1, \dots, n_i$ ). Consider a latent categorical variable  $\mathbf{c}$  with  $K$  classes,  $\mathbf{c}_i = (c_{i1}, c_{i2}, \dots, c_{iK})$ , such that  $c_{ik} = 1$  if individual  $i$  belongs to class  $k$  and zero otherwise. For individual  $i$  in latent class  $k$  at time point  $t_{il}$  ( $l = 1, \dots, n_i$ ) we have e.g.:

$$Y_{il} = \alpha^k + a_i^k + \beta^k t_{il} + b_i^k t_{il} + \varepsilon_{il}^k.$$

The class specific fixed effects  $\alpha^k$  and  $\beta^k$  describe the mean profiles of the  $k^{th}$  group. For ease of notation, possible group-specific effects of covariates have been suppressed in the above expression of the mean profiles. The random effects  $a_i^k$  and  $b_i^k$  follow a bivariate normal distribution  $N(0, D^k)$ , and  $\boldsymbol{\varepsilon}_i^k = (\varepsilon_{i1}^k, \dots, \varepsilon_{iT}^k)' \sim N(0, \Sigma_i^k)$ . The growth curves across different latent sub-populations may differ in terms of the mean profile, measurement error, and variance-covariance matrix of the random effects. In general, the latent categorical variable  $\mathbf{c}$  relates to covariates  $\mathbf{x}_i$  via a multinomial logistic regression model:

$$P(c_{ik} = 1 | \mathbf{x}_i) = \frac{e^{\delta_0^k + \boldsymbol{\delta}_1^k \mathbf{x}_i}}{\sum_{k=1}^K e^{\delta_0^k + \boldsymbol{\delta}_1^k \mathbf{x}_i}},$$

with  $\mathbf{x}_i$  containing covariate information for subject  $i$ .

Parameter estimates for this model can be obtained using maximum likelihood with an EM algorithm (Dempster, Laird and Rubin, 1977).

Grouping the individuals into the  $K$  classes is done via the posterior distribution of the random effect, conditional on  $y_i$ . An individual is assigned to the class for which it has the highest posterior probability.

When, for each individual, more than a single outcome is measured over time, a multivariate set of longitudinal profiles is obtained. Interest could be in finding subgroups of individuals that are similar in their evolution over time for the various repeated sequences. Then, the goal is to find clusters that are unique in the evolutions over time of the different outcomes, as well as in the correlation structure over time and between these outcomes. In particular, a model to analyze the evolution of two related outcomes was presented by Nagin and Tremblay (2001) and by Nagin (2005). Let  $\mathbf{Y}_1 = (Y_{11}, Y_{12}, \dots, Y_{1T_1})$  and  $\mathbf{Y}_2 = (Y_{21}, Y_{22}, \dots, Y_{2T_2})$  denote two longitudinal profiles to be modeled together. The dual trajectory model assumes that the  $J$  trajectory groups of  $\mathbf{Y}_1$  are probabilistically linked to the  $K$  groups for  $\mathbf{Y}_2$ . In addition, the model makes the assumption of conditional independence given group membership (as is the case in the single trajectory group-based approach).

Conceptually, the extension of the dual model to more than two outcomes is straightforward. From a practical point, the addition of outcomes results in an unmanageable proliferation of probability matrices linking the trajectories for the various outcomes. Still, there are many circumstances where it is valuable to link trajectories of three or more outcomes of interest.

Applications of the multitrajectory modelling approach can be found in Nagin (2005) and Piquero et al. (2002).

Extensions of the group-based approach for multivariate longitudinal data to the growth mixture modeling setting is even more problematic, stemming from the high dimension of the joint distribution of the random effects (see Section 3). An alternative two-stage approach for two outcomes is described by Putter et al. (2008). In the first stage, a latent class model is estimated based on the longitudinal data of the first outcome. In the second stage, the relation between the latent classes, patient characteristics, and other outcome(s) are studied. Roy and Lin (2000) propose a latent variable model for repeated measures for different outcomes that are assumed to measure an underlying quantity of main interest. They relate the observed outcomes to a latent variable by means of random (intercept) effects regression models. The random intercepts are independent, i.e., conditional independence of the outcomes given the latent variable applies. The latent variable is modeled as a function of covariates by a separate linear mixed model.

In this paper, we propose a clustering algorithm for multivariate longitudinal data. The method resembles a  $k$ -means iterative clustering procedure. The idea is to divide the observations in  $k$  clusters such that the full likelihood for the  $m$  jointly measured repeated outcomes becomes maximized. Therefore, a joint-modelling approach using mixed models is implemented. Fieuws and Verbeke (2008) use this pairwise modeling strategy in a discriminant analysis. They predict renal graft failure by fitting bivariate mixed models to 4 repeatedly measured markers. The obtained estimates were used in Bayes rule to obtain the prognosis for long-term success of the transplant, at each point in time.

Before specifying the different steps of the clustering algorithm, some background on fitting a joint model for multivariate longitudinal data is given in the next section.

### 3 Joint Model for Multivariate Longitudinal Data

Denote for a univariate longitudinally measured outcome  $Y_{il}$  the  $l^{th}$  measurement available for the  $i^{th}$  subject ( $i=1, \dots, N$ ;  $l=1, \dots, n_i$ ), and let  $\mathbf{Y}_i$  denote the vector of all measurements for the  $i^{th}$  subject. Linear mixed models assume that the vector  $\mathbf{Y}_i$  satisfies:

$$\mathbf{Y}_i | \mathbf{b}_i \sim N(\mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i, \Sigma_i).$$

Here,  $\mathbf{X}_i$  and  $\mathbf{Z}_i$  are  $(n_i \times p)$  and  $(n_i \times q)$ -dimensional matrices of known covariates,  $\boldsymbol{\beta}$  is a  $p$ -dimensional vector of regression parameters, called the fixed effects,  $\mathbf{b}_i$  is the  $q$ -dimensional vector containing the subject-specific or random effects assumed to be sampled from the  $q$ -dimensional normal distribution with mean zero and covariance  $D$ , and  $\Sigma_i$  is a  $(n_i \times n_i)$  covariance matrix which depends on  $i$  only through its dimension  $n_i$ . Marginally,  $\mathbf{Y}_i$  follows a normal distribution with mean  $\mathbf{X}_i \boldsymbol{\beta}$  and covariance matrix  $V_i = \mathbf{Z}_i D \mathbf{Z}_i' + \Sigma_i$ .

When  $m$  longitudinally measured outcomes are available, a joint-modelling approach using mixed models can be followed. The full (log)likelihood

$$\sum_{i=1}^N \ell_i(\mathbf{Y}_{1i}, \mathbf{Y}_{2i}, \dots, \mathbf{Y}_{mi} | \Theta^*) \quad (3.1)$$



has to be maximized. In this expression,  $l_i$  is the loglikelihood contribution of subject  $i$  to the full joint mixed model and  $N$  is the number of subjects. This full joint model can be specified as random-effects models for each outcome process, and the processes are linked by imposing a joint multivariate distribution on the random effects. Although this approach has many advantages, its usability is limited by the dimension ( $m$ ) of the data. In case the number of repeated outcomes becomes large, computational problems are likely in the estimation process due to the high dimension of the joint distribution of the random effects.

Instead of maximizing the likelihood of the full joint model a pairwise approach can be used to obtain estimates for all parameters in the full joint model. Pseudo-likelihood estimation (Besag, 1975), replaces the joint full likelihood by a suitable product of marginal or conditional densities, where this product is easier to evaluate than the original likelihood. Fitting all possible pairwise models is equivalent to maximizing a pseudo-likelihood function of the following form:

$$\sum_{i=1}^N \sum_{(r,s)} l_{rsi}(Y_{ri}, Y_{si} | \Theta_{rs}),$$

$r = 1, \dots, m-1$ ;  $s = r+1, \dots, m$ ;  $l_{rsi}$  the log likelihood contribution of subject  $i$  in the bivariate model for outcomes  $r$  and  $s$ ; and  $N$  is the total number of subjects.  $\Theta_{rs}$  represents the vector of all parameters in the bivariate joint mixed model corresponding to the specific pair  $(r, s)$  of repeated outcomes. Let  $\Theta$  be the stacked vector combining all pair-specific parameter vectors  $\Theta_{rs}$ . There are  $m(m-1)/2$  bivariate joint models to be considered. Some parameters in  $\Theta^*$  (in 3.1) have a single counterpart in  $\Theta$ , e.g., the covariance between random effects from two

different outcomes. Other elements in  $\Theta^*$  have multiple counterparts in  $\Theta$ , e.g., the covariance between random effects from the same outcome. Given that the pairwise approach fits within the pseudo-likelihood framework, an asymptotic multivariate normal distribution for  $\Theta$  can be derived. Asymptotic normality of the pseudo-likelihood estimator in the single parameter case and in the vector-valued parameter case is shown in Arnold and Strauss (1991) and in Geys et al. (1999). Finally, estimates for the parameters in  $\Theta^*$  can be calculated by taking averages over all pairs. This is obtained by

$$\bar{\Theta}^* = A\bar{\Theta} \sim N(\Theta^*, A\Sigma(\bar{\Theta})A')$$

with  $A$  a matrix containing the appropriate coefficients to calculate the averages and  $\Sigma\bar{\Theta}$  equals the covariance matrix for  $\bar{\Theta}$  obtained by an expression shown in Arnold and Strauss (1999).

A mean estimate is simply obtained by averaging all corresponding pair-specific estimates in  $\bar{\Theta}$ . Standard errors of these estimates take into account the variability amongst the pair-specific estimates. Furthermore, estimates corresponding to two pairwise models with a common outcome are based on overlapping information and hence are correlated. This correlation is also accounted for in the sampling variability of the combined estimates in  $\bar{\Theta}^*$ .

The idea of replacing the full likelihood by pairwise marginal likelihoods is used in Step 3 of the proposed clustering algorithm.

## 4 Clustering Algorithm

In this section, we propose an algorithm to reveal latent subgroups (also referred to as clusters, classes or components) for multivariate repeated outcomes. The idea is that the data are not coming from one multivariate distribution, but rather that the generation process behind the data is a mixture of a number of multivariate normal distributions, each described by a density  $N(\boldsymbol{\mu}^k, \Sigma^k)$ . The algorithm aims to locate the  $N$  observations in a number of clusters such that the pseudo likelihood, based on pairwise models of this multivariate normal distribution is maximal. The algorithm is iterative in nature and resembles a partition clustering method. The cluster criterion being used is the individual pseudo-loglikelihood contribution. An illustration of the algorithm is discussed in Section 5 and the performance of the algorithm is investigated in Section 6. The algorithm consists of the following steps:

1. Choose the number of clusters,  $k$ .
2. Randomly divide the  $N$  observations into  $k$  initial clusters.
3. Iterate the following steps until no observation switches cluster anymore:
  - (a) Fit all bivariate joint models with the  $k$  clusters as ‘known’ groups (see Section 3). For each bivariate joint model, based on outcomes  $r$  and  $s$ , this results in  $k$  mean profiles over time,  $\boldsymbol{\mu}_{rs}^k$  as well as  $k$  covariance matrices  $\Sigma_{rs}^k$ .
  - (b) For each pair  $p = (r, s)$  of longitudinal outcomes ( $p = 1, \dots, P$ ) the following  $k$  likelihoods for observation  $i$  are then calculated:

$$L_{pi}^k = L_{rs,i}^k = (2\pi)^{-n_i/2} |\Sigma_{rs}^k|^{-1/2} e^{-\frac{1}{2} (\mathbf{y}_{rsi} - \boldsymbol{\mu}_{rs}^k)' (\Sigma_{rs}^k)^{-1} (\mathbf{y}_{rsi} - \boldsymbol{\mu}_{rs}^k)}.$$

(c) The sum of natural logarithms of these likelihoods over all  $P$  pairs is a natural choice as a cluster criterion.

$$\ell_i^k = \sum_{p=1}^P \log(L_{pi}^k).$$

For each observation  $i$  this results in  $k$  individual pseudo-likelihood contributions. Each observations is (re-)classified into the group having the highest individual pseudo-likelihood. Steps (a)–(c) are repeated until no observations change cluster.

We will dwell on Steps 1 and 2 in turn.

#### 4.1 Optimal Number of Clusters

Deciding on the optimal number of clusters is a difficult task, which has not yet been completely resolved. Often, the problem of determining the optimal value for  $k$  is separated from estimating the parameters of the cluster-specific distributions for a fixed  $k$ . The approach most often used in applied research to estimate the number of clusters is based on penalized information criteria. The likelihood increases with the addition of components to the model and therefore needs to be corrected by a term to penalize for model complexity (the number of estimated parameters in the model). A wide variety of such criteria exists, but the Akaike information criterion (AIC, Akaike, 1974) and the Bayesian information criterion (BIC, Schwarz, 1978) are among the ones most commonly used. The performance of the AIC and BIC criterion have been investigated in the

model-based clustering (via mixtures) context. AIC tends to overestimate the correct number of clusters (Soromenho, 1993; Celeux and Sormenho, 1996). Fraley and Raftery (1998) note that there is considerable support for the use of BIC in this situation.

Counterparts for the AIC and the BIC information criteria for model selection were derived for the framework of pseudo-likelihood. We refer to Varin and Vidoni (2005) for a derivation of the pseudo AIC criterion, and to Gao and Song (2010) for the pseudo BIC criterion. The criteria have the usual form, but the effective number of parameters is to be estimated from the sensitivity matrix,  $H(\theta)$ , and the sandwich estimator,  $G(\theta) = H(\theta)J^{-1}(\theta)H(\theta)$ , with  $J(\theta)$  the variability matrix of the pseudo score functions. The effective number of parameters is  $\dim(\theta) = \text{tr}(H(\theta)G^{-1}(\theta))$  and

$$\begin{aligned} AIC &= -2p\ell(\bar{\theta}_{pl}, y) + 2\dim(\theta), \\ BIC &= -2p\ell(\bar{\theta}_{pl}, y) + \log(N)\dim(\theta), \end{aligned}$$

where  $p\ell$  is the pseudo-likelihood function evaluate at  $\bar{\theta}_{pl}$ . In mixed-effects models, the observations in general are not independent. Therefore, the effective sample ( $N_e$ ) size replaces the sample size in BIC and can vary from the number of subjects  $N$  to the number of observations,  $N \times m \times n$ , for balanced measurements. The effective sample size can be obtained from the correlation matrix (Faes, 2009).

Thus, to determine the optimal number of clusters it is advisable to perform the clustering algorithm for a range of  $k$ -values (e.g.,  $k = 2, 3, 4, \dots$ ) and evaluate the need to include an

additional cluster in the model by means of the Akaike and Bayesian information criterion. The optimal number of clusters is found when the information criterion reaches a minimum.

## 4.2 Random Division in $k$ Initial Clusters

It is well known that the results of a partition clustering method can depend on the starting cluster seeds, both in the number of clusters found and in their centroids. An unfortunate choice of the division to start the procedure in the first step could lead to a poor final division of the data into  $k$  groups. To minimize this risk, it is advisable to repeat the proposed cluster algorithm for a number of times, with different randomly chosen  $k$  initial groups. Out of these runs, the replicate giving the highest value for the pseudo-likelihood value,  $\sum_{i=1}^N \ell_i^k$  has to be selected as the final solution. This strategy reduces the possibility of accepting a poor solution due to an inappropriate starting seed. In addition, reproducibility of the pseudo-likelihood value is an indication of how well a particular number of groups fits the natural structure of the data.

## 5 Application to EEG Data

To illustrate the proposed cluster algorithm, the method was applied to an Electro-Encephalogram (EEG) study conducted at Janssen Pharmaceutica (Belgium).

### 5.1 Introduction to the Data

The aim of EEG studies is to characterize the effects of psychotropic drugs on cortical brain activity, on the basis of spectral electro-encephalograms. Depending on the frequency measurements range, this activity is referred to as delta activity (below 4 Hz per second), theta

activity (4–7.5 Hz per second), alpha activity (8–12.5 Hz per second), beta activity (13–30 Hz per second), and gamma activity (above 30 Hz per second). The activity is refined in low and high activity (e.g.,  $\alpha_1, \alpha_2, \dots$ ).

Delta activity is normally seen in babies or in adults in slow-wave sleep. Theta activity is seen in children or during drowsiness or arousal in adults. Alpha waves are seen when a person is alert in a relaxed way, closing the eyes. Beta activity (low amplitude) is often associated with active, busy, or anxious thinking and active concentration. Rhythmic beta waves are linked with pathological or drug-related causes. Gamma waves are related with strong mental activity like solving problems, fear, and awareness.

The data being analyzed includes 10 psychoactive agents at 4 different doses, including a placebo dose. To each compound, 32 rats were randomly assigned, 8 per dose group. Although the brain waves of rats and humans are observed in comparable frequency bands, not all functionalities are the same. There are, however, more similarities than differences, making experiments measuring the electrical brain activity in rats very interesting to study the effect of psychoactive agents on the activity of human brains. The compounds included in the study are: (1) Psychostimulants: Amphetamine, Nicotine; (2) Antidepressant: Bupropion; (3) Cholinesterase inhibitors: Donepezil, Galantamine, Tacrine; (4) Anti-epileptics: Lamotrigine, Valproate; and (5) NMDA receptor antagonists: Memantine, PCP. Cholinesterase inhibitors are used to treat moderate to severe dementia of the Alzheimer's type. The anti-epileptics listed are used in the treatment of mania. The NMDA receptor antagonists are used for different purposes. Memantine is used to treat moderate and severe dementia of the Alzheimer's type and in that view could be

listed with the cholinesterase inhibitors. PCP in low to moderate doses acts as a stimulant, whilst at higher doses it has a sedative effect.

Forty-five minutes after administration of the psychoactive agent, the brain signals of the rats in active wake state are monitored every 15 minutes during 1.5 hours, at six different positions in the brain (left and right frontal, left and right parietal, left and right occipital). For each rat 9 activity profiles are obtained, at the 6 different positions in the brains. To illustrate the clustering algorithm, we focus on the frequencies obtained at the left prefrontal cortex. So we are facing 9-variate longitudinal profiles. To be able to compare the results with analyses done in the past, we only include the placebo and the highest dose level. This reduces the data set to 160 rats of which 139 have follow-up data. That said, very comparable results were obtained when including all four dose levels in the analyses (data not shown).

To visualize the data, the individual longitudinal profiles are given in Figure 1 for the 9 frequency measurements. The response of interest is the percentage change with respect to the measurement at baseline  $Y_{ib}$  (administration of the drug):  $Y'_{ij} = 100(Y_{ij} - Y_{ib}) / Y_{ib}$ . At baseline all percentage changes are by definition equal to zero. The graphical display therefore excludes the baseline data. In graphical displays and in the statistical models, time zero refers to the first measurement obtained after administering the drug (i.e., after 45 minutes). Heterogeneity is seen in all waves; some rats have a decrease in the frequency while for others an increase is obtained as an effect of the drug. For some waves extreme profiles are seen, such as for the  $\alpha_1$  wave. This heterogeneity is of course caused by administering 10 different drugs at different dose levels. When applying the clustering algorithm, this information was not taken into account. The goal of



the analyses is to see if it is possible to identify subpopulations within the set of 139 rats. Subpopulations that are homogeneous in the growth parameters for the 9 waves and in the correlation structure. The information about the compounds and doses will later be used to assess whether the identified groups are meaningful.

## 5.2 Applying the Clustering Algorithm

The cluster analysis method was applied for the 9-variate response profile by fitting 36 bivariate joint models as explained in Section 4. Models specifying 1, 2, 3, 4, 5, 6, 8, and 10 clusters were fitted to the data. Generally, one names the frequency measurement ranges, obtained for each rat, as alpha, beta, delta, and gamma activity. We will reserve these Greek letters to refer to the fixed or random effects in our statistical model and use  $Y_{m,il}$  for the percentage change of the  $m^{th}$  frequency measurement, obtained for rat  $i$  at time  $t_{il}$  with  $t_{i1} = 0$  referring to the first follow-up measurement taken 45 minutes after administration of the drug.

For each longitudinal profile ( $m = 1, \dots, 9$ ) the following random-intercept model was specified ( $k = 1, \dots, K$ ;  $t_{i1} = 0, \dots, t_{i8} = 7$ ):

$$Y_{m,il} = \alpha_m^k + a_{mi}^k + \beta_m^k t_{il} + \gamma_m^k t_{il}^2 + \varepsilon_{m,il}^k.$$

The parameters  $\alpha_m^k$ ,  $\beta_m^k$ , and  $\gamma_m^k$  describe the average quadratic evolution of outcome  $m$  over time. For each of the  $K$  clusters a separate mean trajectory is fitted. Note that in this model, the cluster is incorporated as a known group effect. To initiate the algorithm this grouping is obtained by randomly dividing the observations in  $K$  initial clusters.

The random intercept  $a_{mi}^k$  takes into account heterogeneity within cluster  $k$  and introduces correlation between the measurements of response  $m$  over time. Associations between the nine longitudinal profiles were imposed by assuming that the random intercepts  $a_{mi}^k$  and  $a_{m'i}^k$  are distributed as a bivariate normal distribution with mean zero and a  $2 \times 2$  covariance matrix. The errors  $\varepsilon_{m,il}^k$  are zero mean normally distributed with variance  $\sigma_{\varepsilon_m}^2$ . More specifically,

$$\begin{pmatrix} a_{mi}^k \\ a_{m'i}^k \end{pmatrix} \sim N \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{a_m}^2 & \sigma_{a_{mm'}}^k \\ \sigma_{a_{mm'}}^k & \sigma_{a_{m'}}^2 \end{pmatrix} \right].$$

The random-effect and error distributions can be specified as cluster specific, allowing the associations between the nine longitudinal response profiles to differ from cluster to cluster. At maximum, 8 rats are randomized to each dose-compound combination. Cluster-specific covariance matrices for the multivariate normal distribution, estimated via the pseudo-likelihood estimation method, turned out to be singular. This may signal a perfect dependency among some variables, induced by an overspecified model fitted to a small set of data. In the application, we therefore assume the random effects and error distributions to be common to all clusters. This model results in constant correlation over time between measurements of the same response, and between measurements of different responses.

### 5.3 Number of Clusters

The AIC and BIC value for pseudo-likelihood estimation were obtained for models imposing  $K = 1, 2, 3, 4, 5, 6, 8, 10$  clusters. To start the algorithm, the 139 rats were randomly divided

into  $K$  groups. To minimize the risk of choosing unfortunate starting values the algorithm was executed 35 times for the same number of clusters specified. The run resulting in the highest pseudo-likelihood value, based on the pairwise model described in Section 3 is reported here. Table 1 shows the pseudo-likelihood value, the AIC and BIC value, and the effective number of parameters for the different models. A sample size of 139 was used to obtain the BIC. These criteria balance the increase in pseudo-likelihood value with the increase in model complexity. A graphical display is presented in Figure 2. It is seen that the BIC value gradually decreases from one to five/six components, from where the BIC appears to level off. The AIC value still decreases when adding components to the model; but the AIC is known for overestimating the number of clusters in data (Hu and Xu, 2003). In what follows, we will discuss the results for the model specifying 5 clusters. The pseudo-likelihood values could be duplicated for the settings with 2–5 clusters (the maximum value is obtained for respectively 20, 4, 4, and 6 runs out of the 35). For the models specifying more than five clusters the maximum pseudo-likelihood value could not be duplicated. For the setting with six clusters a number of runs result in comparable pseudo-likelihood values (minus twice pseudo-log-likelihood values: 617.64; 618.14; 618.19; and 618.25).

#### 5.4 Composition of the Clusters

Figure 3 graphically displays the composition of the clusters resulting from the model specifying 5 clusters; it is seen that the algorithm results in a natural grouping of the different doses and compounds included in the study.

The clustering algorithm grouped all placebo dose levels of the psychoactive compounds and the active dose of Bupropion in one cluster (cluster 3:  $n = 81$ ). It is not unexpected that the placebo dose levels of the compounds are found in one group. But it is interesting to see that the effect of Bupropion on the brain activity, as quantified by the 9 frequency measurements, cannot be distinguished from the effect of a placebo level. In the model specifying 8 clusters, the rats on Bupropion are separated from this cluster.

The second to largest cluster (cluster 2:  $n = 22$ ) contains the rats that were randomized to the highest dose level of the cholinesterase inhibitors (Donepezil, Galantamine, Tacrine,) or Nicotine. In the model specifying 8 clusters, the rats on Nicotine are separated from this cluster.

The rats randomized to the active dose of Memantine or PCP are grouped together, both psychoactive compounds are NMDA receptor antagonists (cluster 5:  $n = 18$ ).

Cluster 1 ( $n = 11$ ) groups the rats randomized to the active dose level of the anti-epileptics (Lamotrigine, Valproate). The clustering algorithm does separate the rats on the highest dose of amphetamine from the rest (cluster 4:  $n = 7$ ). Indicating that the two psycho-stimulants included in the study (amphetamine and nicotine) exhibit different effects on the brains.

## 5.5 Multivariate Longitudinal Profiles

The estimated mean growth trajectories for each cluster and for each frequency measurement are displayed in Figure 4. For the placebo dose level cluster, i.e., cluster 3, Figure 4 shows flat wave profiles at a mean percentage change value close to zero. No effect of the psychoactive compounds is noticed on the rats brain activity.

Cluster 4, i.e., the highest dose level of Amphetamine, is characterized by its profiles for the  $\delta$ ,  $\theta_1$ , and  $\alpha_1$  waves. Amphetamine is the only psychoactive compound in the study that results in a reduction in the  $\delta$  and  $\theta_1$  frequencies and an increase in  $\alpha_1$  frequencies. The induced change for  $\alpha_1$  vanishes by the end of the study.

The effect of administering the highest dose of Memantine or PCP (cluster 5) is best seen in the  $\beta_1$ ,  $\beta_2$ ,  $\alpha_2$  and  $\gamma_2$  waves. For this cluster, the reduction in frequency measurement for the  $\beta_1$ ,  $\beta_2$ , and  $\alpha_2$  waves is larger than observed for the other clusters. The  $\gamma_2$  frequencies increased. This increase is larger than seen for the other clusters and the effect is still present by the end of the study.

Cholinesterase inhibitors and nicotine at the highest dose (cluster 2) result in distinct profiles for  $\theta_1$ ,  $\alpha_1$ ,  $\alpha_2$ , and  $\gamma_1$ . A positive percentage change is observed for  $\theta_1$  and  $\gamma_1$ , this increase seems to level off around time point 5. The percentage changes of  $\alpha$  waves are negative during the study.

The anti-epileptic compounds (cluster 1) behave different for the  $\delta$ ,  $\beta_2$ ,  $\theta_2$ , and  $\gamma_2$  frequencies. The percentage change for the  $\delta$  frequencies is positive and more or less constant during the study, the  $\beta_2$  percentage change is also positive but keeps increasing during the study period.

These results show that rats randomized to the same compounds were nicely clustered together. The compounds constituting a cluster are known to give rise to similar effects on cortical brain activity, as measured by EEG. Thus, the results can be interpreted from a clinical point of view.

## 6 Performance of the Algorithm: Simulations

In this section, we explore the performance of the proposed clustering algorithm by means of simulated labelled data. Settings representing separated and overlapping clusters, of varying cluster sizes (equal and unequal sizes) are considered. The performance of the algorithm will be examined in terms of the choice of the optimal number of clusters and in terms of the correct grouping of the observations.

### 6.1 Data Generation Model

Multivariate longitudinal data for four clusters (indexed by  $k$ ), 6 profiles (indexed by  $m$ ), at 7 timepoints (indexed by  $t_{il}$ ) is generated according to the following random-intercepts model for a linear evolution with time:

$$Y_{m,il} = \alpha_m^k + \alpha_{mi}^k + \beta_m^k t_{il} + \varepsilon_{m,il}^k, \quad (6.2)$$

where  $m = 1, \dots, 6$ ,  $t_{il} = 0, \dots, t_{i7} = 6$ , and  $k = 1, \dots, 4$ .

Values for the cluster by profile specific fixed effects  $(\alpha_m^k, \beta_m^k)$  are given in Table 2. The random intercepts  $\alpha_{mi}^k$  and random error terms  $\varepsilon_{m,il}^k$  are generated from multivariate normal distributions, common to all clusters. Association between the 6 profiles is introduced by allowing the random intercepts to co-vary. Their correlation matrix is given in Table 3.

Data representing different degrees of overlap among the clusters is obtained by increasing the random intercept and residual variances of the 6 profiles. Table 4 displays the considered settings for the variance terms and the corresponding average Mahalanobis distance between the (centers

of the) four clusters. For setting 1, the random intercept variances were specified to equal (2.0, 0.5, 3.0, 2.0, 2.0, 1.0) and the residual variances were fixed to (1.0, 0.3, 1.0, 1.0, 2.5, 1.0). Settings 2 — 9 are obtained by multiplying these variances by a factor 1.0, 1.5 or 2.0. The generated data consists of four clusters with smallest Mahalanobis distances between clusters (1,3) and (2,4). These clusters are thus harder to separate from each other. The cluster sizes were specified to be equal to each other and equal to 7, 10, and 15 in turn. For settings 1, 5, and 9 data was also generated for unequal cluster sizes:  $(n_1 = 20, n_2 = 10, n_3 = 5, n_4 = 5)$  and  $(n_1 = 40, n_2 = 20, n_3 = 10, n_4 = 10)$ . Although the unequal cluster size setting is very limited and purely illustrative, it can shed some light on the performance of the clustering algorithm when facing unequal cluster sizes.

For each setting, 50 data sets were generated. Figure 5 shows the longitudinal profiles obtained for setting 1 and assuming equal cluster sizes of 10. Clearly, the four clusters cannot be discerned by naked eye.

## 6.2 Performance of the algorithm: Results

The clustering algorithm was entertained for  $k = 1, 3, 4$ , and 5, with 15 random initial divisions and a maximum of 45 iterations for step 3 of the algorithm. Each iteration involves 15 bi-variate mixed models. For each response, the model given in (6.2) was applied to the simulated data. The two random intercepts were allowed to co-vary. The variance-covariance matrices for the random intercepts and residual errors were specified to be common to the clusters. Due to increasing computing time with increasing values for  $k$  and  $n_k$ , the algorithm was not entertained for larger values of  $k$ .

In order to evaluate the performance of the clustering algorithm in choosing the optimal number of clusters the AIC and BIC values for the pseudo-likelihood estimations were obtained. The change in information criteria with increasing model complexity (.i.e. increasing value of  $k$ ) is graphically displayed in Figure 6. Although the IC values gradually decrease when adding components to the model, they appear to level off at  $k = 4$ . This indicates that a good balance between model fit and model complexity is reached for four clusters.

Table 5 displays the mean (and standard deviation), minimum and maximum value of the distribution of correctly classified observations when the algorithm is used with four clusters. In fact the algorithm allocates each observation into one of four groups. To circumnavigate the label degeneracy all  $4!$  permutations of the labels, assigned by the algorithm, have to be compared with the true label. Table 5 presents the results for the permutation maximally recovering the true labels. For example, for data generated under setting 2 with clusters of size 7, we conclude that on average the clustering algorithm is able to assign 84% of the observations to the proper cluster.

As expected, we see that the larger the overlap between the clusters, as measured by the Mahalanobis distances between their centers, the harder it is to reveal the grouping in the data (Table 5). The four sub-populations are well recovered by the algorithm for the settings 1, 2, 4 and 7. Also for settings 3, 5 and 8 the proportion of correctly classified observations is reasonably good. For settings 1, 5, and 9 data was simulated for a total of 40 observations under equal and unequal cluster sizes. Comparison of the performance of the clustering algorithm for



both situations seems to indicate that the performance is slightly weaker when clusters have different sizes.

## 7 Discussion

In this paper, we presented an algorithm to reveal clusters in the setting of multivariate repeated data. The algorithm mimics a  $k$ -means algorithm. The means of the  $k$  clusters are obtained via bivariate joint models for the repeated responses. An individuals' pseudo-likelihood contribution is used as the criterion to classify an individual into a cluster.

The algorithm was applied on EEG rat data for 9 longitudinally measured frequencies; with clustering performed at rat level. It turns out that rats randomized to the same compounds were nicely clustered together. The compounds constituting a cluster are known to give rise to similar effects on cortical brain activity, as measured by EEG. Thus, the results could be interpreted from a clinical point of view. The proposed algorithm is an exploratory tool that has potential value to divide a heterogeneous population in homogeneous subpopulations. The performance of the algorithm was also investigated by means of simulations. For the settings considered, the algorithm turns out to have a good performance, both in terms of determining the optimal number of clusters and in terms of classifying the observations in the correct cluster.

The longitudinally measured wave responses in the EEG study were all continuous and recorded at the same time points. The proposed algorithm is, however, not confined to this type of data. The cluster algorithm breaks down to repeatedly fitting bivariate mixed models. As such, it benefits from the attractive feature of a multivariate mixed model that different responses (e.g.,

binary and continuous) can be combined; and to sequential responses. For incomplete responses, missingness at random (MAR) is assumed within each pair of repeated responses. If this assumption does not hold, the pairwise mixed models do not give valid parameter estimates and the results of the clustering algorithm could fail to be trustworthy. Study design features, such as, for example, blocking, can be incorporated in the clustering algorithm, by introducing another random effect in the bivariate model.

The primary practical limitation of the algorithm is computation time. Given that the algorithm is iterative in nature, computation time increases with increasing complexity of the data set; i.e., increasing number of individuals, number of repeated outcome measures and number of specified clusters. The BIC criterion was used to choose the optimal number of clusters. The bootstrap likelihood ratio test is an alternative (McLachlan, 1987), but was not implemented because of computing time.

In general, cluster analyses are sensitive to starting values and to outliers. This is not different for the proposed algorithm. Running the algorithm for a number of times, each time starting from a different random division in  $k$  initial groups, allows to evaluate sensitivity to starting values. The maximum pseudo-likelihood value is harder to duplicate when more cluster are specified; again increasing computing time.

The issue of outliers is harder to investigate. Outliers induce clusters with a few units, centered around the outlier. Bivariate mixed models, as many other statistical tools, easily run into problems when applied to small and sparse data sets (cluster).

Model misspecification is a concern in any statistical modeling exercise. Misspecification of the cluster specific distributions could, for example, result in an overestimation of the number of clusters present in the observations. However, a detailed study of the effect of model misspecification on the results of the clustering algorithm, by means of simulations, is out of the scope of this article.

## References

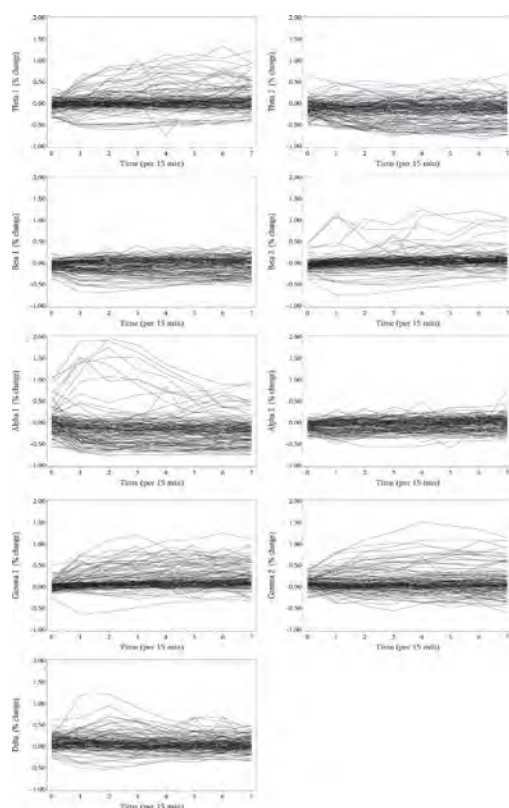
- [1] Arnold, B. C. and Strauss, D. (1991). Pseudolikelihood estimation: some examples. *Sankhya B.*, **53**, 233–243.
- [2] Besag, J. (1975). Statistical Analysis of Non-Lattice Data. *The Statistician*, **24**, 179–195.
- [3] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.
- [4] Faes, C., Molenberghs, G. Aerts, M., Verbeke, G., and Kenward M (2009). The Effective Sample Size and an Alternative Small-Sample Degrees-of-Freedom Method. *The American Statistician*, **63**, 389–399.
- [5] Fieuws, S. and Verbeke, G. (2008). Predicting Renal Graft Failure Using Multivariate Longitudinal Profiles. *Biostatistics*, **9**, 419–431.
- [6] Gao, X. and Song, P. X. K. (2010). Composite likelihood Bayesian information criteria for model selection in high-dimensional data. *Journal of the American Statistical Association*, **105**, 1531–1540.

- [7] Geys, H., Molenberghs, G., and Ryan, L. (1999). Pseudo-likelihood modelling of multivariate outcomes in developmental toxicology. *Journal of the American Statistical Association*, **94**, 34–745.
- [8] Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, **61**, 215–231.
- [9] Hu, X. and Xu, L. (2003). A Comparative Study of Several Cluster Number Selection Criteria. *Intelligent Data Engineering and Automated Learning Lecture Notes in Computer Science*, **2690**, 195–202.
- [10] Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*, Pearson Prentice Hall.
- [11] Jones, R. (2011). Bayesian information criterion for longitudinal and clustered data. *Statistics in Medicine*, **30**, 3050–3056.
- [12] Land, K. C. and Nagin, D. S. (1996). Micromodels of Criminal Careers: A Synthesis of the Criminal Careers and Life Course Approaches via Semiparametric Mixed Poisson Regression Models, with Empirical Models. *Journal of Quantitative Criminology*, **12**, 163–191.
- [13] Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**, 963–974.
- [14] MacQueen, J. B. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the 5th Berkely Symposium on Mathematical Statistics and Probability* **1**, 281–297

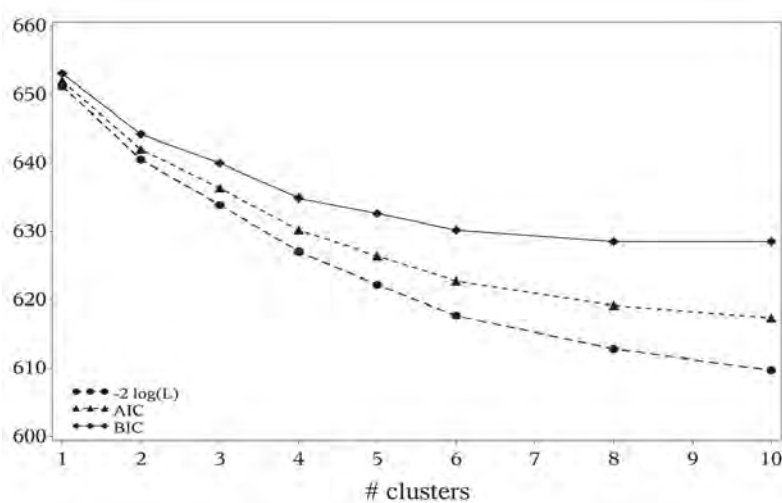
- [15] McLachlan G. (1987). On Bootstrapping the Likelihood Ratio Test Statistic for the Number of Components in a Normal Mixture. *Journal of the Royal Statistical Society. Series C*, **36**, 318–324
- [16] McLachlan G. and Basford K. (1988). *Mixture models, Inference and Applications to Clustering*. Marcel Dekker, Inc.
- [17] McLachlan G. and Peel D. (2000). *Finite mixture models*. New York: Wiley.
- [18] Muthén, L. K. and Muthén, B. O. (1998-2007). *Mplus User's Guide. Fourth edition*. Los Angeles, CA: Muthén and Muthén.
- [19] Muthén, B. O. and Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM-algorithm. *Biometrics*, **55**, 463–469.
- [20] Nagin, D. S. and Land, K. C. (1993). Age, criminal careers, and population heterogeneity: specification and estimation of a nonparametric, mixed Poisson model. *Criminology*, **31**, 327–362.
- [21] Nagin, D. S. (1999). Analyzing developmental trajectories: a semiparametric, group-bases approach. *Psychological Methods*, **4**, 139–157.
- [22] Nagin, D. S. and Tremblay, R. E. (2001). Analyzing developmental trajectories of distinct but related behaviours: a group-based method. *Psychological Methods*, **6**, 18–34.
- [23] Nagin, D. S. (2005). *Group-based Modeling of Development*. Cambridge, MA.: Harvard University Press.

- [24] Nylund, K. L., Asparouhov, T., and Muthén, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: An Interdisciplinary Journal*, **14**, 535–569.
- [25] Piquero, A. R., Brame R., Mazzerole P., and Haaplanen R. (2002). Crime in Emerging Adulthood. *Criminology*, **40**, 137–70.
- [26] Putter H., Vos T., de Haes H., and van Houwelingen H. (2008). Joint analysis of multiple longitudinal outcomes: Application of a latent class model. *Statistic in Medicine*, **27**, 6228–6249.
- [27] Roy, J. and Lin, X. (2000). Latent variable models for longitudinal data with multiple continuous outcomes. *Biometrics*, **56**, 1047–1054.
- [28] Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.
- [29] Varin, C. and Vidoni, P. (2005), A note on composite likelihood inference and model selection. *Biometrika*, **92**, 519–528.
- [30] Verbeke, G., and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.

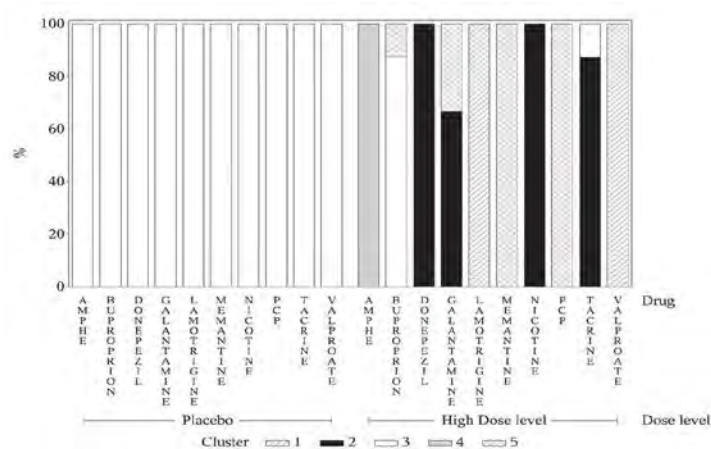
**Figure 1:** Rat specific profiles for the 9 repeatedly measured frequencies (time zero is 45 min after baseline) - EEG Study.



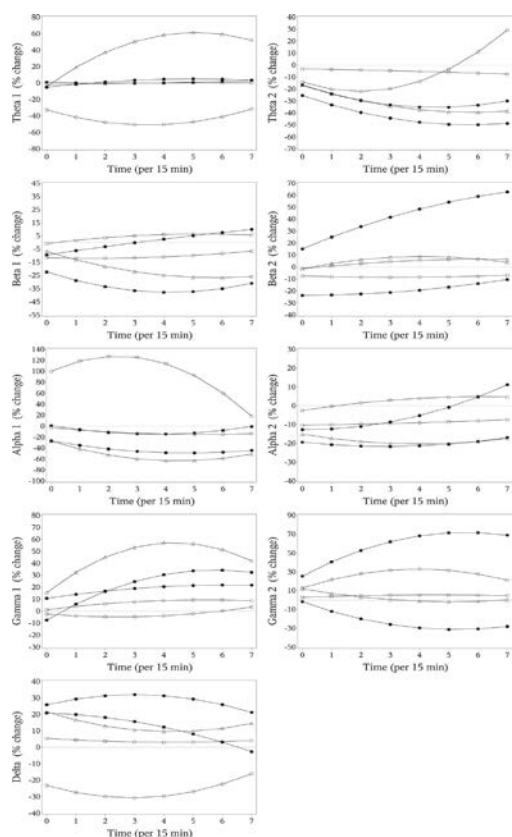
**Figure 2:** Minus twice pseudo-log-likelihood, AIC, and BIC for the different models - EEG Study.



**Figure 3:** Composition of the clusters in terms of psycho-active compound and dose - EEG Study.

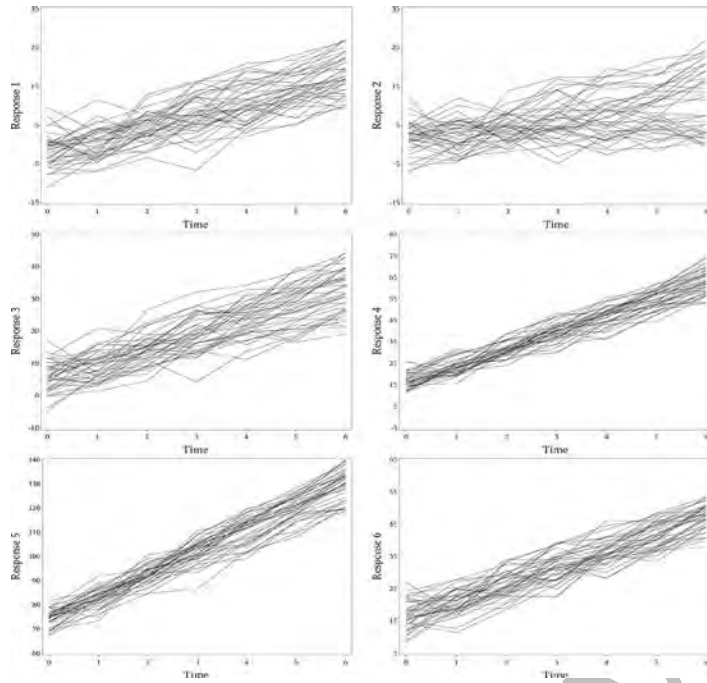


**Figure 4:** Estimated cluster mean profiles for the 9 repeatedly measured frequencies (time zero is 45 min after baseline) - EEG Study.

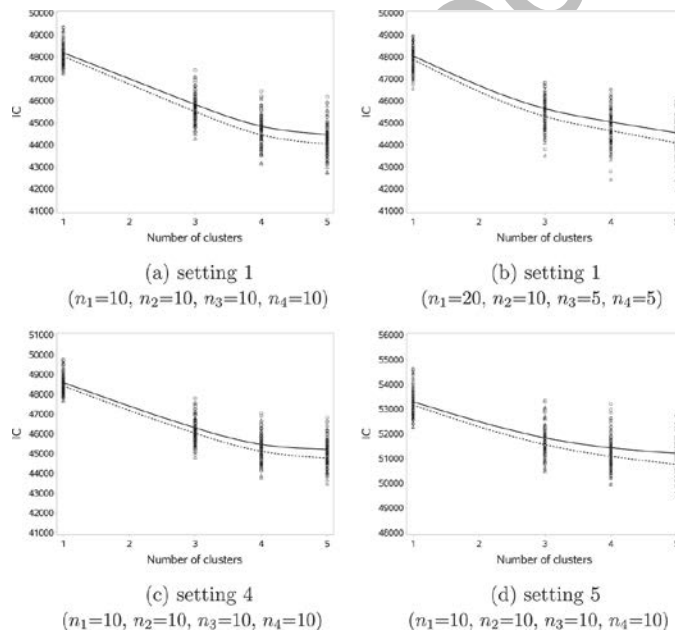




**Figure 5:** Simulated multivariate longitudinal data for four clusters. Data is generated according to Model (6.2), setting 1 (Tables 2 and 4) and cluster sizes equal to 10.



**Figure 6:** AIC (dashed lines,  $\Delta$ ) and BIC (full lines,  $\circ$ ) for the models resulting from the clustering algorithm for the multivariate longitudinal data generated according to Model (6.2) for settings 1, 4 and 5 (Tables 2 and 4).



**Table 1:** Minus twice pseudo-loglikelihood values and information criteria for the final models resulting from the clustering algorithm - EEG Study.

# clusters	effective # of parameters	-2pl	AIC	BIC
1	366	651260	651992	653065
2	755	640466	636317	639923
3	1229	633859	636317	639923
4	1581	627028	630190	634829
5	2108	622189	626406	632592
6	2550	617623	622723	630205
8	3179	612800	619157	628485
10	3808	609677	617293	628467

**Table 2:** Values specified for the fixed intercept and slope effects to simulate data according to Model (6.2).

Cluster ( $k$ )	Response ( $m$ )	$\alpha_m^k$	$\beta_m^k$	Cluster ( $k$ )	Response ( $m$ )	$\alpha_m^k$	$\beta_m^k$
1	1	-3	3	3	1	-3	4
1	2	3	0	3	2	4	0
1	3	5	6	3	3	5	4
1	4	17	7	3	4	17	8
1	5	74	8	3	5	74	10
1	6	13	6	3	6	20	5
2	1	2	3	4	1	-3	3
2	2	3	3	4	2	-3	3
2	3	6	5	4	3	5	3

2	4	17	7	4	4	16	8.5
2	5	74	9.5	4	5	75	9
2	6	15	5	4	6	19	4

**Table 3:** *Values specified for the correlation matrix of the random intercepts to simulate data according to Model (6.2).*

1.00	0.25	0.10	0.30	0.20	0.00
	1.00	0.20	0.10	0.10	0.10
		1.00	0.10	0.10	0.10
			1.00	0.20	0.10
				1.00	0.10
					1.00

**Table 4:** Settings considered to simulate data according to Model (6.2): multiplying factors for the random intercept variances (2.0, 0.5, 3.0, 2.0, 2.0, 1.0) and residual error variances (1.0, 0.3, 1.0, 1.0, 2.5, 1.0) in Model (6.2) and average Mahalanobis distance between the four clusters.

	Multiplying factors for Random intercept variances	Multiplying factors for Residual error variances	Average Mahalanobis distance
Setting 1	1	1	20.3
Setting 2	1	1.5	17.1
Setting 3	1	2	15.2
Setting 4	1.5	1	19.4
Setting 5	1.5	1.5	16.2
Setting 6	1.5	2	14.2
Setting 7	2	1	19.0
Setting 8	2	1.5	15.7

Setting 9	2	2	13.8
-----------	---	---	------

**Table 5:** Proportion of correctly classified observations (average over 50 simulations (standard deviation), minimum (min) and maximum value (max)) under the settings specified in Table 2 and Table 4 - Multivariate longitudinal data generated according to Model (6.2).

Setting	Average Mahalanobis Distance	$n_k = 7$			$n_k = 10$			$n_k = 15$		
		mean (std)	min	max	mean (std)	min	max	mean (std)	min	max
1	20.3	99 (4)	79	100	100 (1)	93	100	100 (1)	98	100
2	17.1	84 (15)	46	100	81 (14)	55	100	89 (14)	47	100
3	15.2	65 (14)	43	93	61 (15)	35	98	56 (12)	30	82
4	19.4	97 (6)	75	100	99 (4)	75	100	100 (1)	95	100
5	16.2	75 (14)	46	100	75 (16)	43	100	76 (16)	47	100

6	14.2	59 (11)	39	86	53 (12)	40	93	50 (9)	32	70
7	19.0	96 (8)	71	100	98 (5)	78	100	100 (1)	95	100
8	15.7	71 (3)	43	96	70 (14)	48	100	68 (14)	45	100
9	13.8	54 (9)	39	75	54 (9)	35	85	48 (9)	32	77

Setting	Average Mahalanobis Distance	$n_1 = 20, n_2 = 10, n_3 = 5, n_4 = 5$			$n_1 = 40, n_2 = 20, n_3 = 10, n_4 = 10$					
		mean (std)	min	max	mean (std)	min	max			
1	20.3	85 (15)	63	100	96 (11)	64	100			
5	16.2	64 (11)	48	90	62 (10)	41	88			
9	13.8	53 (7)	38	70	49 (5)	38	60			